

# The Validity of Health Risk Appraisal Instruments for Assessing Coronary Heart Disease Risk

KEVIN W. SMITH, MA, SONJA M. MCKINLAY, PhD, AND BRUCE D. THORINGTON, BA

**Abstract:** This study evaluated the validity of the scoring systems employed by 41 health risk assessment instruments (HRAs) with respect to the probability of death due to coronary heart disease. Validity was assessed by comparing predictions of mortality risk produced by each HRA to estimates from the Framingham Heart Study and the Risk Factor Update Project. Correlations with both epidemiologic estimates indicated that instruments employing logistic regression or the Geller/Gesner methodology had the highest

validity coefficients, while validity was lowest for self-administered general health status and lifestyle questionnaires. However, most instruments using the Geller/Gesner technique appear to systematically overestimate the probability of CHD mortality. For HRAs based on additive risk scales, validity was often attenuated by the crude categorization of some risk factors and by the omission of the effects of age from the scoring system. (*Am J Public Health* 1987; 77:419-424.)

## Introduction

In the past decade, there has been a dramatic increase in the use of health risk appraisal instruments (HRAs).<sup>1</sup> HRAs utilize information about such risk indicators as blood pressure, weight, smoking habits, and medical history to predict the likelihood of morbidity or mortality due to various causes. These instruments appear to have two primary uses. First, many organizations are using HRAs to characterize the general health status of their client or employee populations and thus inform policies with respect to facilities and benefit plans. Second, because they focus attention on the lifestyle factors that influence risk, HRAs have also been widely advocated as a means to alert individuals to their personal health risks, usually in conjunction with some lifestyle modification program.<sup>2</sup>

A 1982 editorial in this Journal<sup>3</sup> reviewed this burgeoning industry and made a strong plea for more rigorous epidemiological and statistical input to improve the quality of HRAs. As a first step toward this goal, this paper summarizes the findings of an assessment of the relative validity of the cardiovascular risk scoring systems employed by existing HRAs. Underlying this analysis is the assumption that if an HRA does not have a valid method for deriving risk scores, then there is little point in conducting further assessments of the instrument's worth. The paper focuses on the risk of Coronary Heart Disease, the outcome most often appraised by HRAs.

## Methods

Consistent with standard psychometric usage, the validity of a research instrument may be defined as the degree to which the instrument measures the outcome it was designed to measure.<sup>4</sup> It is important to distinguish validity from HRA effectiveness. Validity is a matter of measurement accuracy. Effectiveness, on the other hand, refers to an HRA's ability to elicit desired behavioral changes. The effectiveness of an HRA may have little to do with its validity as a measurement instrument.

In this study, mortality due to coronary heart disease (CHD) during a 10-year period was chosen as the most appropriate outcome for assessing the validity of health risk appraisal instruments. This outcome was selected for several reasons. First, heart disease is the leading cause of death for American adults. Second, all of the HRAs reviewed included heart disease risk factors and most provided separate risk estimates for CHD. Moreover, CHD mortality is comparatively easy to define, heart disease data for various populations are widely available, and risk indicators for this outcome are well established in the literature.

We decided that the most appropriate criteria for inferring CHD mortality risk were the estimates available from large epidemiologic studies. Since the effects of specific CHD risk factors tend to differ somewhat from study to study, models from two different investigations were employed as criteria for assessing instrument validity.

The first set of criterion models was derived from the experiences of participants in the Framingham Heart Study. Several restrictions were imposed to identify Framingham subjects who were appropriate for modeling risk. Biennial examination periods were evaluated with respect to sample sizes, age group distributions, the availability of physiological measures, and the frequency of CHD deaths. This review led to the selection of the fourth biennial examination (administered predominantly in 1956) as the baseline period with mortality status at the ninth examination (1966) serving as the 10-year follow-up point. All persons dying from causes other than cardiovascular diseases during the follow-up interval and those known to have heart disease at baseline were eliminated from the sampling frame. A total of 3,604 cases in the Framingham data base (1,564 men and 2,040 women) were found to meet these sampling criteria.

Prior to estimating the models, a small subsample of test cases was randomly selected from the sampling frame. These test cases, representing the characteristics of White American adults, were set aside for later use in correlation analyses. Consideration of the precision of correlation coefficients suggested that sample sizes in excess of 100 would be sufficient to reliably identify instruments that explained at least half of the variation in heart disease risk. Therefore, samples of 120 men and 120 women were randomly selected from the data base and reserved for use as test cases.

The remaining 3,364 cases were used to estimate the Framingham criterion models of heart disease mortality. In this group, 3 per cent of the men and 0.8 per cent of the women died of coronary heart disease (as defined in Shurtleff<sup>5</sup>) during the 10-year follow-up period. Logistic

Address reprint requests to Kevin W. Smith, MA, Research Scientist, Cambridge Research Center, American Institutes for Research, 1100 Massachusetts Avenue, 3rd floor, Cambridge, MA 02138. Dr. McKinlay is Principal Research Scientist, and Mr. Thorington is Senior Computer Programmer, also at CRC. This paper, submitted to the Journal July 15, 1986, was revised and accepted for publication November 4, 1986.

**Editor's Note:** See also related editorial p 409 this issue.

regression analysis<sup>6</sup> was employed to estimate the probability of death due to CHD. Since male heart disease mortality rates are two to three times higher than those for women and the effects of specific risk factors appear to differ by sex, separate analyses were performed for each gender.

The second set of criterion models was developed as part of the Risk Factor Update Project (RFUP), a collaborative effort involving the University of California at Los Angeles (UCLA) School of Public Health, General Health, Inc., and the American College of Preventive Medicine, and funded by the Centers for Disease Control. The RFUP reanalyzed and then merged together the results of several major epidemiologic investigations into a single mathematical model for mortality from myocardial infarction.<sup>7</sup> Like the Framingham models, the RFUP also developed logistic regression equations.

Thus, four equations were utilized—separate models for men and women derived from the Framingham and RFUP analyses. The risk factors and coefficients for each model are shown in the Appendix. Each logistic regression equation provides an estimate of the probability that a person with specific characteristics will die of coronary heart disease within the next 10 years. Since most HRAs express mortality risk in terms of the number of expected deaths per 100,000 persons, probabilities were multiplied by 100,000 to convert them into an equivalent measure.

While the Framingham and RFUP data bases have several shortcomings (including small numbers of CVD events, lack of some important risk factors, and changes in treatment during the period of observation), they are among the most comprehensive available and are the most relevant for the test case cohort used in this analysis.

#### **HRA Classification**

The health risk appraisal instruments assessed by the project were obtained from a wide variety of organizations, including federal, state, and local government agencies, universities, and for-profit corporations. To conduct the validity assessments, the characteristics of the 240 test cases drawn from the Framingham data base were inserted in the scoring algorithms used by each HRA. The following characteristics were available for each test case: age, sex, height, weight, diastolic and systolic blood pressure, serum cholesterol, cigarettes smoked per day, smoking history, left ventricular hypertrophy detected by ECG, and glucose intolerance. All test cases were White with no previous history of diabetes or heart disease.

Test case data were processed several different ways. Whenever possible, either the scoring algorithm or the source code for computer routines used by an HRA was obtained to calculate risk scores for each case. In some instances, personal computer software was modified to read the test case data from an existing disk file. Other providers required that the data be sent to them for internal batch processing. Printouts of the results for individual test cases were then returned to the project. Some HRAs utilized factors that had never been asked of Framingham subjects, such as family history of heart disease or exercise levels. These characteristics were set at average levels for all adults to avoid biasing subsequent comparisons with the criterion models.

Health risk appraisal instruments define risk in different ways and use a variety of methodologies for estimating risk. There are, however, many similarities among HRAs in terms of risk measures, scoring methods and processing routines. The 41 instruments included in this study were classified into five major categories of risk outcomes.

### **Type 1 — Mortality Risk per 100,000 Persons**

The largest set of instruments provides estimates of the risk of death due to heart disease per 100,000 persons over a 10-year period. Each of the HRAs in this group employed some variation of the "actuarial" approach to mortality risk originally developed by Geller and Gesner<sup>8</sup> and popularized by Robbins and Hall<sup>9</sup> and Hall and Zwemer.<sup>10</sup> In this method, risk multipliers are established for levels of physiological characteristics and the risk multipliers are either added or multiplied together to form a composite risk factor for an individual. The composite risk factor is then multiplied by the average mortality rate for individuals of the same age, race, and sex to determine the estimated risk per 100,000 persons. The HRAs in this group differ to some extent with respect to the factors considered and the risk multipliers attached to each factor. All require computerized processing to obtain a final risk estimate. Although it is perhaps the most popular scoring method, the Geller/Gesner approach has often been criticized because the manner in which risk multipliers are combined has no basis in probability theory and because the method fails to account for correlated risk factors.<sup>7,11,12</sup>

### **Type 2 — Morbidity Risk per 100,000 Persons**

Two of the instruments produce estimates of eight-year morbidity risk, rather than mortality risk. Both instruments utilize multiple logistic regression equations taken from previous analyses of the Framingham Heart Study and require computer processing.

### **Type 3 — Overall Heart Disease Risk**

A third group of instruments provides estimates of the overall risk of heart attacks or heart disease. Unlike the HRAs above, these instruments defined risk on arbitrary scales ranging from low to high rather than using probabilities. Each of the HRAs in this group assigns point values to the response categories associated with various risk factors, with higher values indicating greater risk. Scale totals are then computed simply by adding together all of the point values. With one exception, these HRAs are self-scored.

### **Type 4 — Life Expectancy**

In the fourth set of instruments, life expectancy in years serves as a surrogate for risk. Using a self-scored worksheet, life expectancy is computed by adjusting current age for a variety of risk factors. Only factors contributing to heart disease were varied for the analyses reported here. The computations required for these HRAs tend to be more complicated than those for other self-scored instruments.

### **Type 5 — General Health Status**

The HRAs in this final category are based on generalized measures of overall health, variously described as lifestyle, stress, or health risk. In this study, factors unrelated to heart disease (e.g., seat belt use) were ignored so that the resulting measures reflected only CHD risk. All of these instruments, which are the simplest to score by hand, produce additive scales ranging from low to high risk.

# Results

To assess the validity of individual health risk appraisal instruments, estimates of risk for each of the 240 test cases were computed for each HRA and for the Framingham and RFUP criterion models. Even though instruments quantify risk in different ways, persons estimated to have a high probability of CHD death by the criterion models should also be found to have a higher than average risk by a valid HRA regardless of the scale used by that HRA to define risk. The Pearson product-moment correlation between criterion model and HRA estimates was therefore adopted as the principal measure of instrument validity in this study. Correlation coefficients were considered to be more powerful indicators of association than nonparametric techniques because risk is a ratio-level measure. Moreover, since no statistical tests were planned, distributional assumptions were not central to this analysis.

Four coefficients were computed for each instrument—correlations with the Framingham and RFUP criterion models for both men and women. Scatterplots of each relationship were visually inspected to ensure that they were not significantly distorted by outliers. The correlations between the two criterion models themselves were .751 for men and .730 for women, indicating that these two models yield similar ratings of mortality risk.

The results of the correlation analyses are summarized in Table 1. To preserve the confidentiality of the HRAs, individual instruments were represented by an identification number. A single identification code (ID number 16) was used for seven different instruments that were all based on the same scoring algorithm. The first column in the Table shows the average of the four correlation coefficients. Ten of the instruments had an average correlation of .75 or greater, which is approximately the size of the correlation one might expect to find among different epidemiologic study estimates. In general, the magnitudes of the four correlations for an HRA tend to be similar regardless of gender or criterion model.

The magnitudes of the validity coefficients are closely related to the estimation procedure employed by an instrument and the manner in which risk is defined. HRAs predicting mortality or morbidity risk had the highest correlations with the criterion model estimates. This is not surprising since the outcome measure (CHD mortality risk per 100,000 persons) for these instruments is the same as that for the criterion models. Comparatively strong relationships were also found for the two HRAs predicting morbidity, which were similar to the criterion models in both estimation procedure (logistic regression) and data source (Framingham Heart Study). The next highest set of correlations was found for instruments predicting overall heart disease risk or life expectancy. The weakest relationships occurred for the general health status and lifestyle assessment HRAs, although it should be noted that this group of instruments did not provide independent assessments of CHD risk.

Additional analyses were conducted separately depending on the nature of the risk measure. For instruments predicting expected mortality risk per 100,000 persons (Type 1), direct comparisons could be made with the criterion model estimates. The magnitude of the actual risk estimate is important because an HRA could have a high correlation with the criterion models and yet consistently overestimate or underestimate the risk level predicted by these models. Descriptive statistics for the test case sample are summarized

**TABLE 1—Summary of HRA Instrument Correlation Coefficients Ranked by Average Correlation**

HRA ID	HRA Type	Mean Correlation*	Men Framingham	Men RFUP	Women Framingham	Women RFUP
5	2	.800	.727	.851	.729	.858
24	1	.786	.714	.864	.685	.834
4	1	.784	.708	.844	.686	.851
2	1	.775	.690	.834	.681	.849
35	1	.767	.778	.830	.649	.783
25	2	.763	.737	.803	.595	.856
33	1	.762	.789	.747	.665	.822
1	1	.759	.783	.747	.662	.822
27	1	.758	.761	.789	.659	.802
34	1	.754	.716	.747	.642	.862
32	1	.744	.708	.779	.608	.836
31	1	.731	.586	.744	.718	.829
28	1	.714	.729	.676	.632	.797
30	1	.710	.750	.625	.628	.801
3	1	.689	.560	.728	.679	.762
26	1	.648	.611	.707	.585	.680
29	3	.632	.634	.714	.535	.628
8	3	.618	.634	.527	.611	.688
6	3	.613	.679	.608	.510	.640
10	3	.581	.579	.476	.625	.633
7	4	.579	.632	.556	.534	.588
12	3	.576	.576	.446	.580	.680
9	3	.555	.593	.514	.493	.611
14	3	.553	.506	.355	.636	.669
13	3	.549	.575	.446	.548	.617
11	4	.523	.582	.443	.481	.575
15	5	.435	.410	.336	.493	.493
19	5	.272	.322	.263	.287	.215
17	4	.267	.297	.283	.308	.177
20	5	.265	.287	.259	.297	.215
18	5	.258	.417	.220	.239	.144
16**	5	.252	.329	.279	.239	.155
22	5	.244	.234	.259	.318	.163
21	5	.162	.149	.267	.117	.111
23	4	.145	.269	.175	.156	-.025

\*The mean of the four criterion model correlations was computed using Fisher's *r* to *z* transformation.

\*\*Six other instruments use the same scoring algorithm as HRA ID No. 16.

## HRA Type Categories

- 1 = CHD mortality risk per 100,000 persons
- 2 = CHD morbidity risk per 100,000 persons
- 3 = Overall CHD risk
- 4 = Life expectancy
- 5 = General health status

in Table 2 for the 14 mortality risk HRAs as well as the Framingham and RFUP models. The mean risk estimates for the two criterion models were similar for men (average risks of slightly more than 3,000/100,000) and women (averages of 504 and 626/100,000). Most of the HRAs, however, had mean and median estimates that were considerably higher than the criterion model averages. In particular, some HRAs predicted risks for certain women that were far in excess of the maximum risks produced by either criterion model. Since all of the instruments involved in these comparisons employed some variation on the Geller/Gesner procedure, it may be that this approach tends to overestimate actual risk even though the relative ranking of risk is similar to both criterion models. Wiley has previously noted the tendency for this technique to inflate estimates of the risk of death from all causes,<sup>13</sup> and Chaves, *et al*, have found that different HRAs provide widely divergent estimates of CHD risk for the same individual.<sup>14</sup>

The comparability of actual risk estimates was also evaluated for the mortality risk instruments. To examine this aspect of validity, the difference between the criterion model and HRA estimate for each test case was calculated and the

TABLE 2—Descriptive Statistics for Instruments Predicting Mortality Risk per 100,000 Persons

Ten-year Coronary Heart Disease Mortality Risk per 100,000 Persons for Test Cases										
	Men					Women				
	Mean	Standard Deviation	Median	Minimum	Maximum	Mean	Standard Deviation	Median	Minimum	Maximum
Framingham	3,395	4,672	1,533	45	21,670	626	1,072	224	4	5,856
RFUP	3,172	3,415	2,346	215	24,524	504	652	239	35	3,789
HRA ID No.										
2	2,503	2,798	1,464	18	13,069	643	942	260	6	4,891
24	2,703	2,879	1,812	64	15,207	1,353	2,076	466	19	10,101
4	3,086	3,205	1,836	40	14,815	947	1,374	393	11	7,182
27	4,380	4,774	2,792	72	24,214	2,016	3,262	528	12	19,679
35	4,703	5,587	2,801	164	28,800	1,953	3,187	568	21	17,600
3	5,244	6,094	2,901	106	28,454	2,328	3,662	1,128	108	19,179
1	4,570	4,878	3,183	93	31,048	2,558	4,401	792	14	26,537
33	4,607	4,902	3,190	91	30,950	2,674	4,536	798	13	26,917
26	4,919	4,222	3,251	159	19,164	1,850	2,363	766	30	10,423
30	4,821	4,895	3,534	106	26,509	2,491	4,466	661	12	30,030
28	5,419	6,030	3,603	61	33,888	2,555	4,345	648	17	25,990
34	4,930	4,748	3,712	134	26,644	2,353	3,878	834	24	27,502
31	5,969	6,333	4,020	14	34,506	1,932	3,104	676	5	15,074
32	5,798	4,801	4,245	488	23,392	1,458	1,461	808	120	6,514

median absolute difference across cases of each gender was determined. The median rather than the mean absolute difference was used to minimize distortions resulting from a few large prediction errors.

Table 3 shows the median absolute differences for each instrument. Three HRAs had differences for the male test cases that were smaller than the median deviation for the two criterion models (930/100,000). The remaining instruments usually had deviations of more than 1,000/100,000. Differences for women were smaller than those for men, but the probability of death due to CHD is also lower for women. For these instruments, nearly all deviations resulted from risk estimates that were higher than those predicted by any of the criterion models. This implies that some of these HRAs combine risk factors in a way that appears to overstate the probability of death. These results were not affected by whether an HRA reported the risk for artery disease or for

coronary, arteriosclerotic, or ischemic heart disease. Thus, differences among these instruments with respect to correlation coefficients and deviations could not be attributed to the way in which risk measures were defined.

One reason for this overestimation bias may lie in the characteristics of the populations from which models are derived. Epidemiologic studies typically exclude persons known to have heart disease from their risk analyses. Instruments using the Geller/Gesner method, on the other hand, base their average age-race-sex CHD risk levels on national mortality data which include individuals with existing cardiovascular problems. The inclusion of these higher risk individuals may therefore explain in part why the Type 1 HRAs yield higher probabilities than the criterion models. It should be noted, however, that many HRAs contain disclaimers warning that the appraisals are not applicable to anyone known to have heart disease.

Because the remaining instruments measured risk in terms of arbitrary scales, their descriptive statistics cannot be meaningfully compared with the criterion models, as in Table 2. However, a comparison of the score distributions for HRAs based on scales (Types 3, 4, and 5) and HRAs providing mortality or morbidity risk estimates (Types 1 and 2) revealed two important differences between these two categories of instruments.

First, as could be expected, the range of potential scores on the arbitrary scales was considerably more restricted than the ranges for the instruments providing risk per 100,000 estimates. The number of unique additive scale values found for all 240 test cases is shown for these HRAs in Table 4. The overall CHD risk and life expectancy instruments (Types 3 and 4) distributed the sample cases over a minimum of 17 different scale points. In contrast, the general health status questionnaires (Type 5) included a maximum of 18 scale values to differentiate risk and a minimum of only three (in the case of one instrument). The importance of this variation in scale values is demonstrated by the correlations with the criterion models (Table 4). All of the instruments with at least 20 scale points had mean correlations with the criterion model risk estimates of at least .5 or greater; in contrast, only

TABLE 3—Median Prediction Errors for Instruments Providing Mortality Risk per 100,000 Estimates

HRA ID	Median Absolute Deviation for:			
	Men		Women	
	Framingham	RFUP	Framingham	RFUP
2	802	762	148	109
24	702	622	380	280
4	916	777	246	162
27	1,216	973	438	336
35	1,356	1,170	426	317
1	1,249	992	594	630
33	1,264	1,069	602	676
30	1,897	1,238	476	399
3	1,384	984	940	794
28	1,717	1,482	528	470
34	1,835	1,308	603	599
26	1,936	1,504	686	556
32	2,377	1,958	692	602
31	3,205	2,435	462	429
Framingham	—	930	—	140
RFUP	930	—	140	—

TABLE 4—Validity Measures for HRA with Additive Scales

HRA ID	HRA Type	(a) Number of Unique Values	(b) Mean Correlation with Criterion Models*	(c) Mean Correlation Adjusted for Age*	(d) Per Cent Improvement in Correlation**
29	3	33	.632	.672	6.4
8	3	26	.618	.671	8.5
6	3	24	.613	.665	8.6
10	3	29	.581	.701	20.5
7	4	69	.579	.634	9.6
12	3	20	.576	.699	21.4
9	3	39	.555	.643	16.0
14	3	20	.553	.710	28.4
13	3	44	.549	.664	20.8
11	4	36	.523	.574	9.8
15	5	9	.435	.635	46.0
19	5	3	.272	.570	109.6
17	4	18	.267	.594	122.6
20	5	9	.265	.578	118.4
18	5	15	.258	.585	126.8
16***	5	5	.252	.575	128.4
22	5	18	.244	.584	138.9
21	5	7	.162	.538	232.6
23	4	17	.145	.588	304.9

\*The mean of the four criterion model correlations was computed using Fisher's  $r$  to  $z$  transformation.

\*\*Per cent improvement =  $(d) = [(c) - (b)]/(b) \times 100$

\*\*\*Six other instruments used the same scoring algorithm as HRA ID No. 16

HRA Type Categories

3 = Overall CHD risk

4 = Life expectancy

5 = General health status

one of the HRAs yielding 18 or fewer scale values had an average correlation as large as .3.

Two factors appear to account for these restricted scale ranges. First, some instruments fail to consider all of the major risk factors for heart disease. Many of the lifestyle questionnaires, for example, do not request information regarding blood pressure or cholesterol levels. Second, the manner in which specific risk factors are measured is often crude. In many cases, variables are represented by only two or three broad categories. The HRAs with more scale points presumably measure risk with greater precision and this, in turn, is reflected in the validity ratings.

A second fundamental difference among the various types of HRAs has to do with the shapes of their risk distributions. The frequency distributions for the mortality and morbidity instruments (Types 1 and 2) were always skewed to the right, peaking at risk levels of around 1,000 per 100,000 persons with tails extending toward the higher probabilities. On the other hand, HRAs based on scales typically exhibited normal distributions, with most cases clustered around an average risk level. The normal distributions seem to occur because many of these instruments do not explicitly incorporate an individual's age into their health risk calculations. This shortcoming means that persons of different ages with similar physiological characteristics will receive similar scores even though their absolute risks may be quite different.

In an effort to improve the comparability of all instruments, therefore, an age-adjusted correlation coefficient was computed. This was done by regressing criterion model estimates on both the raw HRA scale score as well as the age of the test case and then recording the resulting multiple correlation. These adjustments, also shown in Table 4, produced dramatic improvements in the correlations for most

general health and lifestyle HRAs. However, age-adjusted risk estimates from these HRAs are still not as highly correlated with the criterion model estimates as are the estimates from instruments predicting risk per 100,000 persons.

### Discussion

The validity of over 40 health risk appraisal instruments was assessed in terms of the 10-year risk of death due to coronary heart disease. The best of these HRAs explained over 60 per cent of the variation in the risk estimates derived from four criterion models on CHD mortality. At the other extreme, two HRAs had average correlations of less than .2 with the criterion model estimates.

In general, correlations with the criterion models were strongly associated with the way in which risk was defined. Rank ordered by decreasing validity correlations, the risk definition categories were: 1) mortality and morbidity probabilities, 2) overall CHD risk, 3) life expectancy, and 4) general health status. Even though only CHD-related factors were considered, these associations are largely an artifact of the validity outcome: the more closely an instrument's risk measure approximated a 10-year mortality probability, the greater the validity coefficient for that instrument. There are, however, several characteristics of a health risk appraisal instrument that have important implications for the validity of its scoring system.

The most important characteristic is the sophistication of the estimation method. In current epidemiological research, logistic regression analysis is the most widely used technique for estimating probabilities. Both the Framingham and RFUP criterion models were used on logistic equations. One of the two HRAs using the logistic technique had the highest average correlation with four criterion model estimates, in spite of the fact that these instruments were concerned with morbidity rather than mortality risk. While the actuarial approach to risk estimation developed by Geller and Gesner has been criticized on several grounds, instruments using some modification of this approach often produced risk estimates that were within 1,000/100,000 (1 per cent) of the criterion model estimates. However, many of these HRAs exhibited a tendency to consistently overestimate the risk specified by the criterion models. The instruments with the lowest validity correlations were those that utilized an additive weighting method to generate arbitrary risk scales, although many of these did not focus exclusively on CHD.

A second characteristic of HRAs that influences validity is the range of risk estimates an instrument is capable of producing. In general, the greater the number of different risk values generated, the more valid the instrument is likely to be. All of Type 1 HRAs yield a wide range of probability estimates. But several other instruments, particularly the general health status questionnaires, measured certain risk factors only in broad categories and neglected other known determinants of heart disease. As a result, test cases were assigned to fewer than 10 unique scale points for some instruments, thus severely limiting their validity.

A third characteristic that should be considered in this type of assessment is the degree to which an HRA takes a person's age into consideration. Many of the self-scored HRAs either do not modify their risk estimates for age or do not include it directly, conditioning scores on age categories instead. An important additional measure of validity for these instruments, therefore, was the difference between unadjusted and age-adjusted correlation coefficients, which indicated

that lower validity for most of these instruments was largely due to the omission of age from the scoring system.

Several caveats and study limitations deserve to be emphasized. First, for this analysis the probability of death due to coronary heart disease was chosen as the criterion for assessing validity. While this was the most common outcome predicted by HRAs, there may be other more meaningful measures (e.g., life expectancy, appraised age, or quality of life) that should instead be used as a validation standard. Second, this study examined only one of the many causes of death appraised by most instruments. HRA validity with respect to other causes, especially those for which risk factors are less well established than CHD, is a matter for further research. Third, due to the characteristics of the test cases, the results reported here may not be applicable to Blacks or persons aged 35 years or younger. Fourth, the effectiveness of HRAs as a means of stimulating behavioral

change was not addressed by the study. Finally, this secondary analysis does not consider complications posed by actual HRA usage such as the extent to which a person understands the instructions, knows the physiological measures needed to complete the instrument, or can interpret the results. Many people, for example, are unaware of their blood pressure or cholesterol levels. Problems inherent in completing and understanding HRAs may further compromise the validity of these instruments.

#### ACKNOWLEDGMENTS

This project was supported by Grant No. HL32141 from the National Heart, Lung and Blood Institute, National Institutes of Health, Co-Principal Investigators Drs. Sonja M. McKinlay and John B. McKinlay. The authors are indebted to Susan E. Jennings and Mark A. Chaves for their contributions, and to the many developers of health risk appraisal instruments who provided the information that made this project possible.

**APPENDIX**  
**Criterion Model Logistic Regression Equation Coefficients**

	Men		Women	
	Framingham	RFUP	Framingham	RFUP
Constant	-30.80065	(varies with age)	-19.57957	(varied with age)
Age (years)	.31694	.47080	.14216	.12444
Systolic blood pressure (mmHg)	.02810	.02342	.00755	.01787
Serum cholesterol (mg/100 ml)	.04492	.01051	.01586	.00551
ECG abnormality (dummy)	.06172	.72614	1.86140	.41611
Glucose intolerance (dummy)	.89351	1.00300	1.83246	2.32184
Relative weight (per cent)	.04152	—	.00661	—
Cigarettes smoked per day (square root)	.72810	—	2.16780	—
Current smoker (dummy)	—	.96290	—	.25454
Age*Age (quadratic effect)	—	-.00392	—	-.00015
Age*Cholesterol	-.00077	-.00013	—	—
Age*Cigarettes/day (square root)	-.01063	—	-.03597	—
Age*Current smoker (dummy)	—	-.00103	—	—

Dummy variables were coded 0 = absent; 1 = present in the Framingham model, and 1 = absent; 2 = present in the RFUP model. Constant terms in the RFUP models range from -24.72818 to -23.37030 for men and from -19.18336 to -18.07967 for women.

#### REFERENCES

1. Beery W, Schoenbach VJ, Wagner EH, *et al*: Health Risk Appraisal: Methods and Programs, with Annotated Bibliography. National Center for Health Services Research and Health Care Technology Assessment, DHHS Pub. No. (PHS) 86-3396. Washington, DC: Govt Printing Office, June 1986.
2. Vogt TM: Risk assessment and health hazard appraisal. *Ann Rev Public Health* 1981; 2:31-47.
3. Fielding JE: Appraising the health of health risk appraisal. (editorial) *Am J Public Health* 1982; 72:337-340.
4. Nunnally JC: *Psychometric Theory*. New York: McGraw-Hill, 1967.
5. Shurtleff D: Some characteristics related to the incidence of cardiovascular disease and death. The Framingham Study, 16 year follow-up. *In*: Kannel WB, Gordon T: The Framingham Study, an epidemiological investigation of cardiovascular disease. Section 26. Washington, DC: Govt Printing Office, 1970.
6. Kleinbaum DG, Kupper LL, Morgenstern H: *Epidemiological Research: Principles and Quantitative Methods*. Belmont, CA: Lifetime Learning Publications, 1982.
7. Breslow L, Fielding J, Afifi AA, *et al*: Risk Factor Update Project: Final Report. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control, Center for Health Promotion and Education, 1985.
8. Gesner NB: Derivation of risk factors from comparative data. *In*: Robbins LC (ed): *Proceedings of the Seventh Annual Meeting on Prospective Medicine and Health Appraisal*. Indianapolis, IN: Methodist Hospital of Indiana, 1971.
9. Robbins LC, Hall JH: *How to Practice Prospective Medicine*. Indianapolis, IN: Methodist Hospital of Indiana, 1970.
10. Hall JH, Zwemer JD: *Prospective Medicine*. Indianapolis, IN: Methodist Hospital of Indiana, 1979.
11. Schoenbach VJ, Wagner EH, Karon JM: The use of epidemiologic data for personal risk assessment in health hazard/health risk appraisal programs. *J Chronic Dis* 1983; 36:625-638.
12. Wagner EH, Beery WL, Schoenbach VJ, Graham RM: An assessment of health hazard/health risk appraisal. *Am J Public Health* 1982; 72:347-352.
13. Wiley JA: *Predictive risk factors do predict life events*. Bethesda, MD: *Proceedings of the 16th Annual Meeting of the Society of Prospective Medicine*, 1981; 75-79.
14. Chaves MA, Jennings SE, McKinlay SM, McKinlay JB: Cardiovascular risk: differences among health hazard appraisals. Atlanta, GA: *Proceedings of the 20th Annual Meeting of the Society of Prospective Medicine*, 1985:25-27.